**Term Research Project**

**Gunnar Hansen**

**Introduction**

For this term project, I utilized business analytics to gain actionable insights into Hospital-at-Home (H-a-H) programs in the United States. Hospital-at-Home programs provide care services in the comfort of the patient’s own home and have evolved in their proficiency as technology has continued to improve over time.

My research attempted to determine variables that influence the availability of H-a-H programs on a county level. By analyzing the Area Health Resource Files (AHRF) data provided, I determined variable categories from the codebook that may play a role in influencing H-a-H availability. These included measurements of Socioeconomic Factors, Ruralness, Population, Demographic Factors, Hospital Size, and Insurance Coverage levels. Of these variable categories, I decided that Insurance Coverage’s effect on H-a-H would be, out of the group, the hardest to quantify and because of this decided it would be the primary focus of my research.

The primary research question for my report is as follows:

1. **Does Insurance Coverage Rate in a county significantly influence the availability of Hospital-at-Home programs in that same county?**

To control for these other variables, the following secondary research questions were selected:

1. **How does socioeconomic status affect the relationship between insurance coverage and H-a-H availability?**
2. **Are there differences in H-a-H program availability between urban and rural hospitals when controlling for insurance coverage and socioeconomic factors?**
3. **What role does hospital size play in if a hospital offers an H-a-H program in counties with differing insurance rates?**
4. **Do age demographics influence the availability of H-a-H programs in counties with differing insurance rates?**

**Data**

The AHRF is a comprehensive dataset maintained by the United States HRSA (Health Resources and Services Administration) containing information on a variety of healthcare information and population characteristics on a county level across the United States. The files used in this research report were from the year 2022. It is widely used for healthcare related research.

To effectively analyze my research questions, I determined to use the following healthcare and population characteristics from AHRFas variables to help answer my research questions.

* For the primary research question, variables that contributed to an aggregate of health insurance coverage rate by county were utilized. These included “**% Persons <65 with Insurance**”, “**Medicare Enrollees**”, and “**Market Place Enrollees**”. These variables were chosen because they reflect major sources of health insurance coverage across populations.
* For secondary research questions I used the following variables, selected based on their relevance to key influencing factors:
  + **Socioeconomic Status:** Median Income
  + **Ruralness:** RU Continuum Code (1-9, Urban – Rural)
  + **Hospital Size:** Hospital Beds per 1,000 Residents
  + **Demographics:** Veteran Population Percentage and Senior Population Percentage.

Below shown in **Table 1**. are the descriptive statistics for each of the above variables from the AHRF dataset (2022).

**Table 1. Summary Statistics**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Mean | Median | Standard Deviation | Minimum | | Maximum | | N |
| *Senior Pop. %* | 0.19 | 0.19 | 0.05 | 0.05 | 0.58 | | 3141 | |
| *Veteran Pop. %* | 0.07 | 0.07 | 0.02 | 0.01 | 0.34 | | 3141 | |
| *Median Income* | 52794.41 | 50568.00 | 13880.12 | 25385 | 140382 | | 3141 | |
| *Hospital Beds per 1000 Residents* | 2.92 | 1.82 | 4.89 | 0 | 130.73 | | 3141 | |
| *Insurance Coverage Rate (ICR)* | 0.81 | 0.82 | 0.06 | 0.39 | 0.95 | | 3141 | |
| *Population Est.* | 104160.25 | 25776.00 | 333534.25 | 152 | 10105518 | | 3141 | |

**Methods**

**Data Processing:**

* **Data Cleaning:** Data was cleaned using the Power Query Editor in excel. Columns (in the *ahrf2022.csv* file) that contained variables not being used in the research were removed from the data (deleted) prior to data integration.
* **Data Integration:** “*The HaH Hospitals – ISM6404.xlx*” file containing the cities and hospital names of hospitals that had H-a-H programs available was merged with the “*ahrf2022.csv*” around the county name. Prior to merging the queries, the county name was identified in the *“HaH Hospitals – ISM6404.xlx”* excel file by identifying the county name for each city utilizing AI. The merging of the queries was performed using Power Query Editor after the data had been cleaned.
* **Variable Calculations:** 
  + *Hospital Beds per 1,000 Residents:* (Hospital Beds / County Population) \* 1,000
  + *Veteran Population %:*(# of Veterans / County Population) \* 100
  + *Senior Population %:* (# of individuals >65 / County Population) \* 100
* **Variable Categorization:** 
  + *Rural-Urban Continuum Code:*To reduce the amount of instability RU Scores were grouped together from nine into three categories: Urban (1-3), Suburban (4-6), and Rural (7-9).
* **Outlier Detecting/Treatment:** Variables were not removed from the data as high or low outliers could be meaningful to determining H-a-H program availability. The AHRF data source is also reputable, and extremes were determined to be valid.
* **Removed Geographic Values:** Observations from U.S. Territories and other regions outside of the 50 states were excluded since insurance laws differ in these areas and could compromise the integrity of the analysis
* **Missing Values:**Due to an ample number of complete rows of data missing values were removed from the dataset and excluded from the analysis.

**Statistical Analysis:**

* **Logistic Regression Models:** To assess the relationship between the independent variable (Insurance Coverage Rate), confounding variables (% Veterans, % Seniors, Median Income, Population, RU Continuum Code Category, Hospital Beds per 1,000 Residents) and the dependent variable (H-a-H program availability). Due to the presence of extreme correlation between Population & RU Continuum Code two separate regression models were ran. Each model was ran using **JMP Pro’s ‘Fit Model’ function**.
  + **Simple Logistic Model:** The first model was a simple model with Population Size as the IV and H-a-H program availability as the DV.
  + **Multiple Logistic Model:** The second model was a multiple logistic regression model containing ICR and the rest of the confounding variables.
* **Descriptive Statistics:** Calculated using the “real stats” excel add-in for data analysis.

**Data Visualization and Dashboard Creation:**

* **Data Visualizations:** Data Visualizations outside of the Dashboard were constructed using excel.
* **Interactive Dashboard:** Power BI was used to visualize the data into an informative, interactive dashboard.

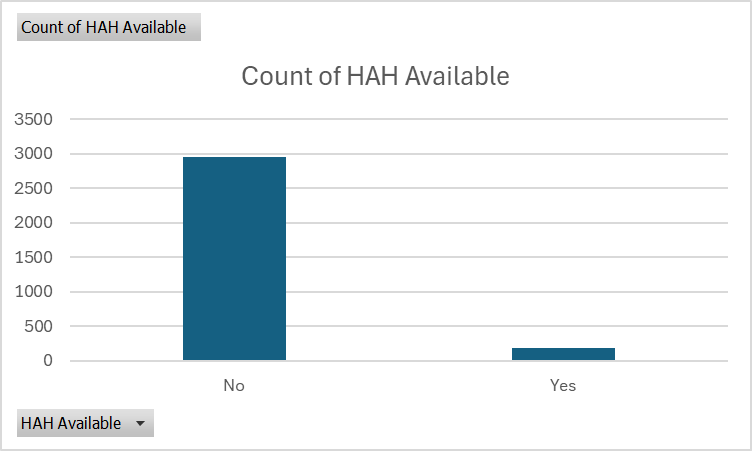
**Data Analysis**

**Key Trends form Exploratory Analysis**

*Counties with H-a-H Programs are far fewer than Counties without them:*

**Figure 1** shows the distributions of H-a-H program availability in the counties studied in the data. From the column chart it is easy to discern counties without a program available far outnumbered counties with programs available. With a smaller minority class (Yes) influencers may stand out more prominently because they drive deviations from the "normal" (majority) pattern.

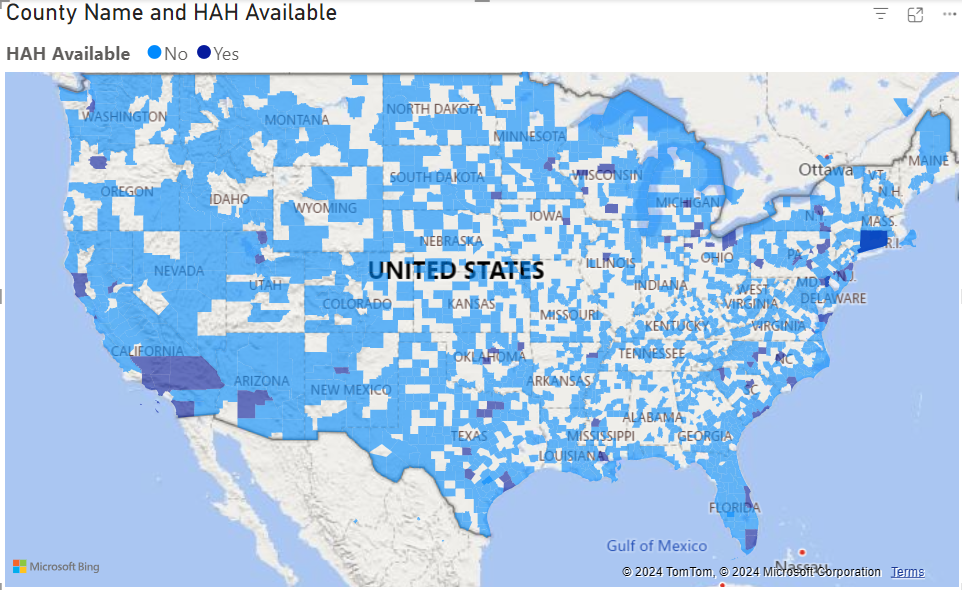
**Figure 1.** H-a-H County Distribution

****

*Geographic Trend Analysis*

The Map visual represented in **Figure 2** below displays counties from the dataset that possess a hospital with a H-a-H program available. There appears to be a strong concentration of H-a-H programs in the northeast and along the Great Lakes, but they are also evenly distributed on the west coast and in Texas. They are noticeably absent around the start of the Rocky Mountains except for Albuquerque, NM and Salt Lake City, Utah H-a-H programs in the map also appear to be clustered in major cities/population centers. The map in **Figure 3** drills a bit deeper showing the insurance coverage rate % by county. From the map we can see that a large portion of Florida in its southwest appears to have poor coverage and analyzing its demographics may be able to offer some valuable insight. Despite this, there does not appear to be any geographical trend for ICR when visualizing it on the county level or any immediate relationship between ICR and H-a-H availability. Ultimately, these visualizations identified that state politics and other regional factors are not significant influencers on H-a-H availability.

**Figure 2.** H-a-H Available Counties Map



**Figure 3.** Insurance Coverage Rate by County

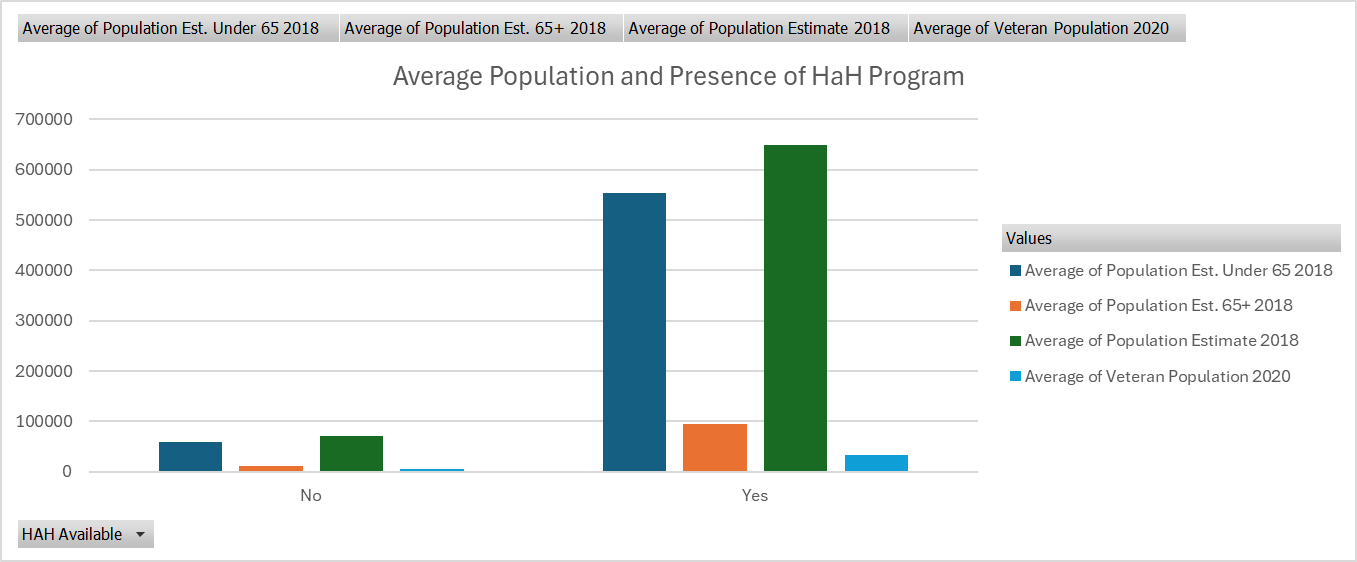
A map of the united states

Description automatically generated

*Trend in Population Size and H-a-H availability.*

An initial trend that can be seen in the data is that there are more H-a-H programs located in larger population centers. **Figure 4** contains a bar chart of average population characteristics for H-a-H available counties vs non-available counties. This chart displayed every population measure included in this research was proportionally higher in areas where an H-a-H is present. Similarly, **Figure 5,** which focuses on ruralness, alludes to how H-a-H program hospitals are primarily located in counties on the lower end of the RU Continuum Code values which indicate a more urban environment, which typically have higher populations. Identifying this trend confirmed population and ruralness may be significant influencers of H-a-H program availability.

**Figure 4.** Average Populations of Counties with H-a-H Presence



**Figure 5.** RU Continuum values and H-a-H availability

A graph of a number of numbers

Description automatically generated with medium confidence

*Trend in H-a-H Availability and the County’s Affluence*

**Figure 6** is a simple column chart displaying the median income in Non-H-a-H counties and counties with H-a-H programs. The chart shows that average median income is significantly higher in areas that have an H-a-H program. This may indicate that H-a-H programs tend to be located in more affluent areas across the country.

**Figure 6.** Median Income in H-a-H Available Areas

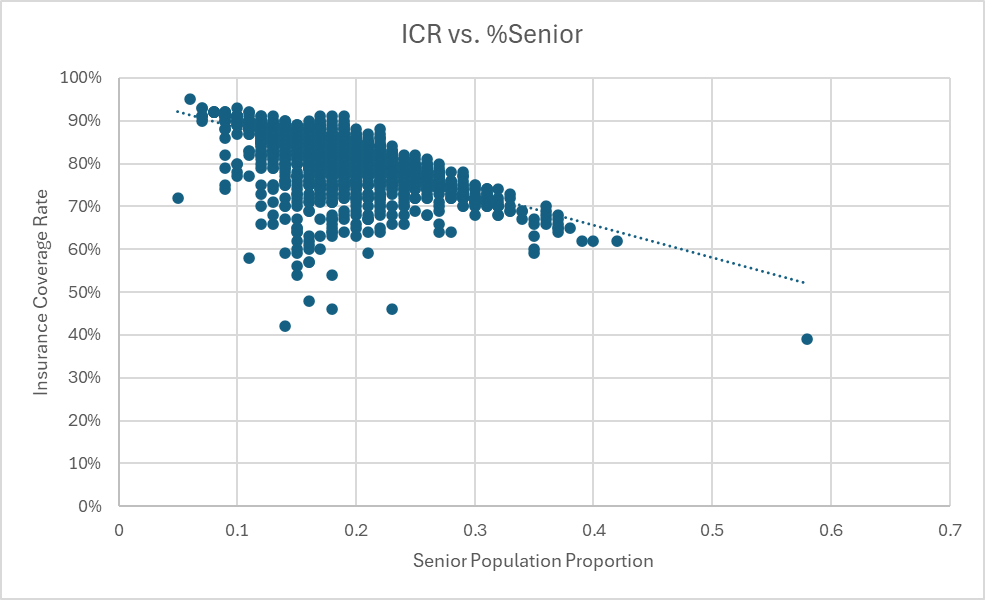
A graph with blue squares

Description automatically generated

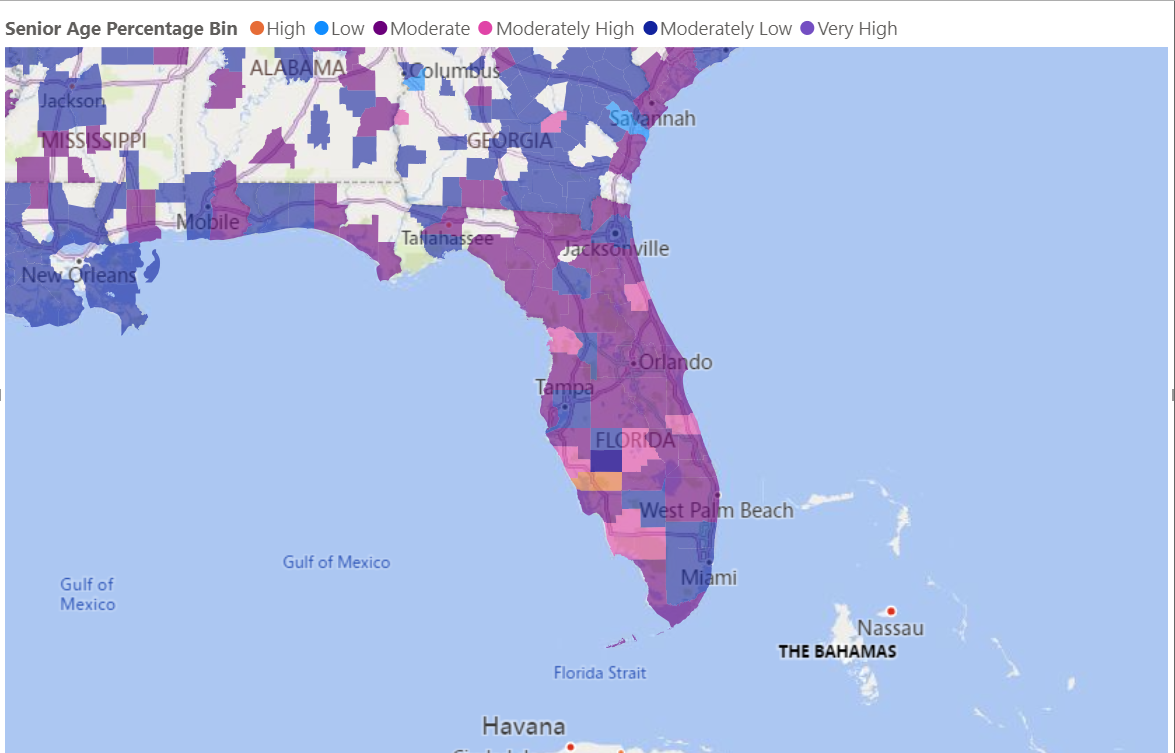
*Trend Between the Percentage of Seniors and Low Insurance Coverage Rate*

Another trend found during exploratory analysis is the negative relationship between Insurance Coverage Rate (ICR) and Senior Population Percentage. **Figure 7** depicts a scatter plot that plots this relationship and illustrates aclear downwards trend showing insurance coverage rates decline as the proportion of seniors in a given county increases. **Figure 8** is a Zoomed-In Map of Florida. As we saw above in the ICR Map Visual (**Figure 3)**, a large portion of SW FL was categorized into low-insurance coverage rate. Looking at the map in **Figure 8**, we can see that this area is also categorized as having a high or moderately high level of seniors amongst the county’s population. This further reinforces that the relationship between these variables could be a significant factor therefore results should taken with a grain of salt as ICR may mask the effect of the % Senior Population and exaggerate its own coefficient or vice versa.

**Figure 7.** ICR %Senior Scatter Plot



**Figure 8.** Florida Zoom-In Senior Population



**Statistical Analysis**

**Population Simple Logistic Model**

The results from the simple logistic regression ran in JMP Pro are illustrated below in **Table 2.** The p-value was less than .0001 and therefore Population was determined as a significant determinant in H-a-H program availability in a county. The r-squared value was found to be .2 indicating the model was not proficient in encompassing all of the significant determinants in H-a-H programs.

**Table 2. Simple Logistic Model Output**

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | **P-Value** | **R-Square** |
| Population Estimate 2018 | <.0001 | 0.2091 |

**Multiple Logistic Regression Model**

**Table 3** shows the corresponding p-values for each of the variables included in the Multiple Logistic Regression Model. For the RU CC Cat. Variable, Urban was used as the reference category for the analysis and the Rural category was deemed significant having a p-value of less than .0001, the Suburban category was not significant having a p-value of 0.8757. Hospital Beds per 1000 residents were deemed significant, also having a p-vale less than .0001. Veteran Population Percentage was deemed to be a significant determinant of H-a-H availability having a p-value of 0.0196. Insurance Coverage Rate and Median Income were also deemed significant with a p-values of 0.0240 & 0.0264 respectfully. Aside from the Suburban category in RU CC Cat, Senior Population Percentage was the only variable not deemed significant with a p-value of 0.7418. Overall, the r-square value for the model was 0.177 which indicates that this model like the simple regression model does not capture a lot of the factors leading to H-a-H availability by county. **Figure 11** displays a column chart of each variable’s logworth.

**Table 3. Multiple Regression P-Values**

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | **P-Value** | **Model R-square** |
| Rural | <.0001 | 0.177 |
| Suburban | 0.8757 | -- |
| Hospital Beds per 1K | <.0001 | -- |
| Veteran Pop. % | 0.0196 | -- |
| ICR | 0.0240 | -- |
| Median Income | 0.0264 | -- |
| Senior Pop. % | 0.7418 | -- |

**Figure 11. Logworth Bar Chart**

A screenshot of a computer

Description automatically generated

**Interpretation of Results**

**Key Insights from Statistical Analyses**

**Overall Model Effectiveness:** From the logistic regression analyses it can be concluded that the simple model, which included only population as a predictor (R² = 0.20), explained a modest 20% of the variance in the outcome, suggesting that population is a key driver. The multiple logistic regression model was similarly effective and had a slightly lower R² of 0.177. Collectively, these results highlight that while population strongly influences the outcome, the modest R² values reflect the complexity of the outcome, which may depend on unmeasured factors not included in the analysis.

**Primary Research Questions**

* **Insurance Coverage Rate:** While ICR significantly predicts H-a-H availability, its relationship with senior population percentages suggests an interplay that warrants further exploration, especially since Senior Population Percentage is a non-significant variable.

**Secondary Research Questions:**

* **Population and Urbanization:** Population was a major influencer on the availability of H-a-H programs in a given county. This indicates that larger, urban populations are strongly associated with H-a-H availability, reflecting resource-driven program distribution.
* **Socioeconomic Status:** Median income was a significant factor, although the weakest of the significant variables. This result leads to the conclusion median income is a critical predictor, underscoring disparities in program access.
* **Geographic Patterns:** The rural-urban divide plays a prominent role in H-a-H availability, with rural counties significantly less likely to offer programs.
* **Veteran Demographics:** The veteran population percentage is a significant factor, highlighting that areas with a higher veteran population are more likely to have H-a-H programs available. This suggests that veteran-specific needs or policies may play a role in the availability of these programs.

**Data Dashboard**

Below **Figure 12** is a screenshot of a data dashboard to explain key findings related to the primary and secondary research questions. Centrally located is information relating to the primary research question and the effect of ICR on H-a-H availability on the county level. Additional visualizations were produced to show the significance of the variables included from the secondary research questions that were deemed significant (demographics, socioeconomic status, ruralness, and population size. Particularly, the ruralness and population sizes of the counties are examined in detail in the dashboard as from the logistic regression outputs, they were far and away the most significant influencers.

**Figure 13. Data Dashboard**

**A screenshot of a map

Description automatically generated**

**Recommendations**

**Recommendation for Future Work**

* **Sub-Group Analysis on Rural Category:** Conducting sub-group analyses within each rural category could provide a more detailed understanding of how factors like insurance coverage rate, veteran population, median income, and hospital beds per 1,000 residents affect H-a-H program availability in each context. This approach can help identify the specific drivers and barriers to H-a-H program implementation that are unique to each county type (rural, suburban, and urban).
  + **Rural:** Conducting a sub-group analysis for rural areas could uncover the impact of unique drivers such as limited access to healthcare infrastructure, transportation challenges, or lower healthcare workforce availability.
  + **Suburban:** For suburban areas, a sub-group analysis might reveal distinct influences, such as their proximity to urban centers and unique variations in socioeconomic diversity.
  + **Urban:** Within urban areas, sub-group analysis could investigate factors like population density, competition between healthcare providers, or greater availability of resources that may uniquely drive or inhibit H-a-H program adoption.
* **Further explore ICR and Senior Population Percentage Interaction:** Further analysis into the interaction of these two terms could reveal whether counties with high senior populations and high insurance coverage rates experience unique programmatic or infrastructural needs, uncover subtle influences not captured in the main model, or identify subgroups (e.g., low-SP%/low-ICR versus high-SP%/high-ICR counties) where targeted interventions could be most effective.

**Recommendations for Improvement**

* **Include Additional Variables:** Future analyses could benefit from adding more granular variables that may influence Hospital-at-Home (H-a-H) availability, such as additional healthcare quality indicators (such as the # of healthcare professionals, the total number of healthcare facilities, the ratio of healthcare providers to patients, the availability of specialized care units, and hospital occupancy rates), additional socioeconomic indicators (creating a composite variable of multiple statistics in addition to Median Income such as the poverty rate, employment rate, educational attainment, and housing affordability), and additional demographic indicator (such as race and ethnicity, gender, and household size) could improve model comprehensiveness. This is reinforced by the fact the regression models only accounted for about 38% of variance in the outcome.